TRABAJO DE FIN DE MÁSTER

Máster Interuniversitario en Estadística e Investigación Operativa UPC-UB

Título: El problema de la separación en modelos de regresión logística

Autor: Óscar Almendros Morón

Director: Klaus Langohr

Departamento: Estadística e Investigación Operativa

Universidad: Universitat Politècnica de Catalunya y Universitat de Barcelona (UPC-UB)

Convocatoria: Octubre de 2018

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

El problema de la separación en modelos de regresión logística

Trabajo de Fin de Máster

Óscar Almendros Morón

Director: Klaus Langohr

Departamento de Estadística e Investigación Operativa

Barcelona, 1 de Octubre de 2018

Agradecimientos

A Klaus Langohr, por su inestimable guía y su inigualable forma de transmitir sus conocimientos sobre la estadística.

A mis compañeros y amigos del máster, por todos esos buenos momentos durante estos dos años. No hace falta que los nombre, ellos saben quienes son.

Y a Míriam, por su paciencia y amor incondicional.

Resumen

Una de las herramientas estadísticas más usadas en el campo de la epidemiología y de la salud pública para analizar el grado de asociación entre una exposición de interés y una enfermedad es la regresión logística, cuya medida de asociación es el *odds ratio*. Este tipo de regresión presenta grandes ventajas, pero también tiene algunos inconvenientes.

Uno de estos inconvenientes es el problema de la separación en los datos. Se trata de un fenómeno que se produce con frecuencia en los modelos de regresión logística y que muestra el grupo de éxitos separado del de los fracasos, es decir, no hay sobreposición entre dichos grupos. Este problema, en caso de que no se tenga en cuenta por el investigador, puede derivar en errores inaceptables, como que impide hallar los estimadores de máxima verosimilitud.

Son varias las posibles soluciones que se propusieron para corregir el problema de la separación, pero la mayoría presentaban limitaciones que ponían en seria duda su uso en la práctica. Fue David Firth, a mediados de los 90, quien desarrolló un método con el que consiguió eliminar completamente este suceso.

En este trabajo se presenta en detalle el problema de la separación y el método de reducción de sesgo. Además se aplica este método a un conjunto de datos de cáncer de endometrio con la intención de comparar los resultados obtenidos con los de otros métodos más tradicionales.

Palabras clave: regresión logística, *odds ratio*, método de máxima verosimilitud, problema de la separación, función *score* modificada, cáncer de endometrio.

Abstract

One of the most commonly used statistical tools in the field of epidemiology and public health to analyze the degree of association between an exposure of interest and a disease is logistic regression, whose association measure is the *odds ratio*. This type of regression has great advantages, but it also has some disadvantages.

One of these disadvantages is the problem of separation in the data. This is a phenomenon that occurs frequently in logistic regression models and shows the group of successes separated from the one of the failures, that is, there is no overlap between these groups. This problem, if it is not taken into account by the investigator, can lead to unacceptable errors, such that it prevents the maximum likelihood estimators.

There are several possible solutions that were proposed to correct the problem of separation, but most had limitations that seriously questioned its use in practice. It was David Firth, in the mid 90s, who developed a method that got to completely eliminate this event.

In this work we present in detail the problem of separation and the bias reduction method. This method is also applied to a set of endometrial cancer data with the intention of comparing the results obtained with those of other more traditional methods.

keywords: logistic regression, *odds ratio*, maximum likelihood method, problem of separation, modified score function, endometrial cancer.

Índice general

Capítulo 1

Introducción

La palabra epidemiología viene derivada del griego *Epi* (Sobre), *Demos* (Pueblo) y *Logos* (Ciencia), lo que nos lleva a poder decir que es aquella parte de la medicina que se dedica a estudiar el desarrollo epidémico y la incidencia de las enfermedades infecciosas en la población.

Esta ciencia tiene como uno de sus propósitos estimar la magnitud o grado de asociación entre una exposición y una enfermedad, pero son muchas las formas que existen de cuantificar esta asociación. Una de las herramientas más utilizadas es el análisis multivariante, donde se encuentra el modelo de regresión logística, que nos ayuda a describir de forma sencilla cómo influye la presencia o no de diversos factores en la probabilidad de aparición de un suceso.

La regresión logística tiene grandes características, como que los términos de error siguen una distribución binomial, que utiliza la función de enlace *logit* o que a la hora de interpretar los parámetros del modelo utiliza el *odds ratio* como medida de asociación. Pero este modelo también presenta algunas limitaciones importantes, como el problema de la separación, un fenómeno por muchos desconocidos y que se produce cuando existe una división entre las variables respuestas y no respuestas. Este suceso se presenta con bastante frecuencia y puede conducir a serios problemas.

Por lo tanto, el objetivo principal de este trabajo fin de máster es explicar a fondo el problema de la separación: desde definir en qué consiste y sus consecuencias hasta ver las posibles soluciones que se han desarrollado a lo largo del tiempo, tanto manual como computacionalente. Para esto último, debemos destacar que hemos trabajado con el software R, un lenguaje de programación que posee un enfoque estadístico.

El contenido de este trabajo se estructurará de la siguiente manera:

Capítulo 2. En este capítulo explicaremos los conceptos más importantes de la epidemiología, los cuales relacionaremos con la regresión logística, de la que hablaremos en profundidad y cuyo modelo tendrá gran importancia a lo largo del trabajo.

Capítulo 3. Este capítulo lo dedicaremos a presentar todo lo relacionado con el problema de la separación, empezando con la introducción de algunos conocimientos importantes sobre la función de máxima verosimilitud y su estimador. Después pasaremos a explicar el problema de la separación en la regresión logística, con las consecuencias que acarrea y las posibles soluciones que se han propuesto para corregir este fenómeno, llegando a el método de reducción de sesgo, propuesto por Firth, el cual describiremos detalladamente ya que es el más utilizado hasta día de hoy. Acabaremos el capítulo viendo un ejemplo empírico donde compararemos este último método con otros para ver su rendimiento.

Capítulo 4. En este capítulo hablaremos sobre lo que es el cáncer de endometrio, viendo sus posibles factores de riesgo y su clasificación, para posteriormente realizar un análisis estadístico con datos sobre este tipo de cáncer donde mostraremos una comparación entre el método presentado en el Capítulo 3 y el método de máxima verosimilitud convencional.

Capítulo 5. Por último, en este capítulo ofreceremos las conclusiones que hemos sacado del trabajo y algunas nuevas propuestas para estudios futuros.

Capítulo 2

Conceptos epidemiológicos

Este capítulo lo dedicaremos a introducir los conceptos y métodos más relevantes sobre la epidemiología, los cuales relacionaremos con el modelo de regresión logística, modelo en el que profundizaremos, ya que es fundamental en el problema de la separación.

Antes de empezar, debemos comentar que buena parte de la notación usada en todo el capítulo se basa en los apuntes de la asignatura de Epidemiología del Máster en Estadística e Investigación Operativa [1], que a su vez fueron sustraídos en su mayoría de *Jewell (2004)* [2].

Comenzamos pues explicando qué es en sí la epidemiología y, aunque son muchas las diferentes definiciones que existen sobre este concepto, destacamos la que se recoge en el diccionario epidemiológico [3], la cual dice:

> "La epidemiología es el estudio de la ocurrencia y distribución de estados o eventos relacionados con la salud en poblaciones específicas, incluyendo el estudio de los determinantes que influyen en dichos estados".

2.1 Diseño de los estudios

Uno de los principales objetivos de la epidemiología es estudiar la relación existente entre una exposición y una enfermedad de interés, y por eso los tipos de estudio que vamos a presentar a continuación serán discutidos de acuerdo con este objetivo.

Existen también distintos criterios a la hora de clasificar los estudios epidemiológicos, pero uno de los más utilizados es el descrito por *Hernández et al. (2000)* [4]. De modo que, según este criterio, podemos destacar cinco tipos de estudios: de ensayos aleatorizados, de cohorte, caso-control, transversales y ecológicos. Sin embargo, los estudios de ensayo aleatorizados y ecológicos se usan con muy poca frecuencia debido a que sus limitaciones muchas veces impiden llevar a cabo dichos estudios, por lo que nos centraremos en los otros tres estudios.

Estudios de cohorte

Son estudios longitudinales que siguen un criterio de selección basado en la exposición de interés, es decir, se toman las muestras entre los individuos expuestos y no expuestos libres de enfermedad. La gran mayoría de ellos son de tipo prospectivo pero se puede dar el caso de que sean retrospectivos.

Estos estudios tienen varias ventajas, como que permiten estudiar más de una enfermedad a la vez y que aseguran que la exposición sea la que provoca la enfermedad, pero también tienen algunos inconvenientes, y es que el tiempo del estudio puede llegar a ser largo y muy costoso e incluso puede haber posibles pérdidas de seguimiento.

Estudios caso-control

Se tratan de estudios que siguen un criterio de selección basado en la enfermedad de interés, de manera que las muestras son tomadas entre los individuos afectados y libres de enfermedad, respectivamente. Son de tipo retrospectivo e ideales para estudiar enfermedades raras.

Como ventajas, permiten el estudio de varios factores de riesgo al mismo tiempo y son bastantes más rápidos y baratos que los estudios de cohorte, pero la gran desventaja es que puede no estar claro si la exposición fue antes de la aparición de la enfermedad, por lo cual es difícil determinar exactamente la exposición. Otra contra de este estudio es el posible sesgo de selección.

Estudios transversales

En estos estudios se toma la muestra de la población de interés y, a diferencia que en los estudios de cohorte, se toma una sola medida del individuo en un tiempo determinado y se observa la presencia o ausencia tanto de la exposición como de la enfermedad.

Entre las ventajas más destacables, son estudios cortos y de bajo coste puesto que no se ha de realizar un seguimiento y pueden servir para generar hipótesis. En cuanto a sus limitaciones, la más grave, al igual que en el estudio caso-control, es que las relaciones causales no pueden establecerse, pues puede no estar claro si la exposición fue anterior o posterior a la enfermedad.

Para entender mejor cada estudio, pasamos a ver la Tabla 2.1, que muestra un resumen de las características por las que nos hemos guiado para poder obtener la clasificación anterior.

Tipos de estudio	Asignación de la exposición	Número de observaciones por individuo	Criterios de selección	Temporabilidad	Unidad de análsis
Ensayo aleatorizado	Aleatoria	Longitudinal	Ninguno	Prospectivo	Individuo
De cohorte	Basada en criterio	Longitudinal	Exposición	Prospectivo o retrospectivo	Individuo
Caso-control	Basada en criterio	Longitudinal o transversal	Enfermedad	Retrospectivo	Individuo
Transversal	Basada en criterio	Transversal	Ninguno	Retrospectivo	Individuo
Ecológico	Basada en criterio	Longitudinal o transversal	Ninguno	Retrospectivo	Población

Tabla 2.1: Clasificación de los estudios epidemiológicos

2.2 Medidas de ocurrencia de una enfermedad

En este punto vamos a hablar sobre la prevalencia y la incidencia, las cuales representan proporciones de caso de enfermedad en una población determinada.

2.2.1 Prevalencia

La prevalencia, P, de una enfermedad es la proporción de individuos afectados por una enfermedad entre la población de interés en un tiempo dado t:

$$P = \frac{X}{N}$$

donde X es el número de casos de enfermedad y N es el tamaño poblacional en el tiempo t.

Pasando a su estimación, sea una muestra de tamaño n, la prevalencia puede ser estimada por:

$$\hat{P} = \frac{X_n}{n}$$

siendo X_n el número de casos entre la muestra.

Dada una muestra de observaciones independientes y asumiendo que X_n sigue una distribución binomial, con parámetros n y P, podemos calcular su intervalo de confianza (CI) usando la distribución binomial o su aproximación a la distribución normal, la cual se basa en el teorema central del límite. De tal forma que:

$$\text{CI}(P; 1-\alpha) = \hat{P} \mp z_{1-\alpha/2}\sqrt{\hat{P}(1-\hat{P})/n}$$

donde $1-\alpha$ es el nivel de confianza y $z_{1-\alpha/2}$ es el $(1-\alpha/2)$-cuantil de la distribución normal estándar.

Un comentario destacable es que la prevalencia se puede calcular en los estudios transversales, pero no en los estudios de cohorte o caso-control.

2.2.2 Incidencia acumulada

La incidencia acumulada, $CI(\Delta)$, de una enfermedad es la proporción de nuevos casos dentro de un período de tiempo de duración Δ entre una población inicialmente libre de enfermedad:

$$CI(\Delta) = \frac{I}{N_0}$$

donde N_0 es el tamaño de la población inicial e I es el número de nuevos casos durante el período de tiempo.

A diferencia de la prevalencia, la incidencia acumulada puede ser calculada en los estudios de cohorte, y su estimador será la proporción de nuevos casos en la cohorte durante el seguimiento. Para calcular su intervalo de confianza, lo podemos hacer igualmente usando la distribución binomial o su aproximación por la distribución normal.

2.3 Medidas de asociación exposición-enfermedad

Pasamos a explicar las medidas para estudiar la asociación entre una exposición y la ocurrencia de una enfermedad.

Lo primero es decir que las medidas que vamos a comentar se han definido para los casos dicotómicos, pero también se puede comparar el riesgo de enfermar entre más de dos grupos de exposición, simplemente eligiendo una categoría de referencia y calculando estas medidas para las otras categorías con respecto a esta.

Otro inciso es que, para agilizar la formulación, a partir de ahora es posible que representemos la exposición mediante la letra E y la enfermedad como D.

2.3.1 Riesgo relativo

El riesgo relativo, RR, es la relación existente entre el riesgo de enfermar en la población expuesta y el riesgo de enfermar en la población no expuesta:

$$RR = \frac{\mathrm{P}(D|E)}{\mathrm{P}(D|\bar{E})}$$

Claramente observamos que el RR es la relación de dos incidencias acumuladas y que nunca puede llegar a ser negativo. Dependiendo del valor resultante se puede interpretar de distinta manera:

- Si $RR > 1$, diremos que existe una probabilidad mayor de D entre las personas expuestas, es decir, E será un posible factor de riesgo para D.

- Si $RR < 1$, E será un posible factor protector para D.
- Si $RR = 1$, indicará que no existe relación entre E y D, lo que significa que son independientes.

Conviene resaltar que el riesgo relativo puede ser estimado solo en los estudios de cohorte, mientras que para los estudios transversales, a la relación de dos incidencias se le llama riesgo relativo de prevalencia, *PRR*. En los estudios caso-control no se puede calcular, ya que es imposible estimar $\mathrm{P}(D|E)$ ni $\mathrm{P}(D|\bar{E})$, por lo que para estos casos usaremos el *odds ratio*, medida de la que hablaremos en breve.

2.3.2 Diferencia de riesgos

La diferencia de riesgos, *RD*, se define como la diferencia entre el riesgo de enfermar en la población expuesta y en la no expuesta:

$$RD = \mathrm{P}(D|E) - \mathrm{P}(D|\bar{E})$$

Esta medida observa la diferencia absoluta, más que la relativa, en los niveles de riesgo, y tiene un rango entre -1 y 1.

En este caso, la *RD* se puede estimar en los estudios de cohorte, pero tampoco es posible estimarlo en los estudios caso-control. En cuanto a los estudios transversales, sería la diferencia entre prevalencias.

2.3.3 *Odds ratio*

El riesgo de enfermar puede igualmente ser expresado mediante *odds*:

$$odds(D) = \frac{\mathrm{P}(D)}{1 - \mathrm{P}(D)}$$

Por lo tanto, el *odds ratio*, *OR*, se puede definir como la relación entre el *odds* de enfermar en los casos expuestos comparados con el *odds* de enfermar en los casos no expuestos:

$$OR = \frac{odds(D|E)}{odds(D|\bar{E})} = \frac{\mathrm{P}(D|E)/(1 - \mathrm{P}(D|E))}{\mathrm{P}(D|\bar{E})/(1 - \mathrm{P}(D|\bar{E}))}$$

Al igual que en el *RR*, y siguiendo la misma numeración, podemos saber a través del *OR* si E es un posible factor de riesgo o protector para D, o si son independientes. Sin embargo, la magnitud del *OR* es más difícil de interpretar intuitivamente, puesto que se trabaja con *odds* y no con probabilidades.

Como ya mencionamos, el *OR* sí que se puede calcular también en estudios caso-control, de tal modo que en estos estudios el *RR* se puede estimar en términos de

odds mediante el OR, siendo $RR \approx OR$ tan grande como $\mathrm{P}(D|E)$ y $\mathrm{P}(D|\bar{E})$ son de pequeños. La relación entre ambas medidas es:

$$OR = RR \cdot \frac{1 - \mathrm{P}(D|\bar{E})}{1 - \mathrm{P}(D|E)}$$

por lo que observamos que el OR está siempre algo más lejos de 1 que el RR, es decir:

$$\left.\begin{array}{l} RR < 1 \\ RR = 1 \\ RR > 1 \end{array}\right\} \Rightarrow \left\{\begin{array}{l} OR < RR < 1 \\ OR = RR = 1 \\ OR > RR > 1 \end{array}\right.$$

De forma gráfica, podemos ver esta propiedad a través de la Figura 2.1, la cual hemos sacado de *Zhang y Yu (1998)* [5].

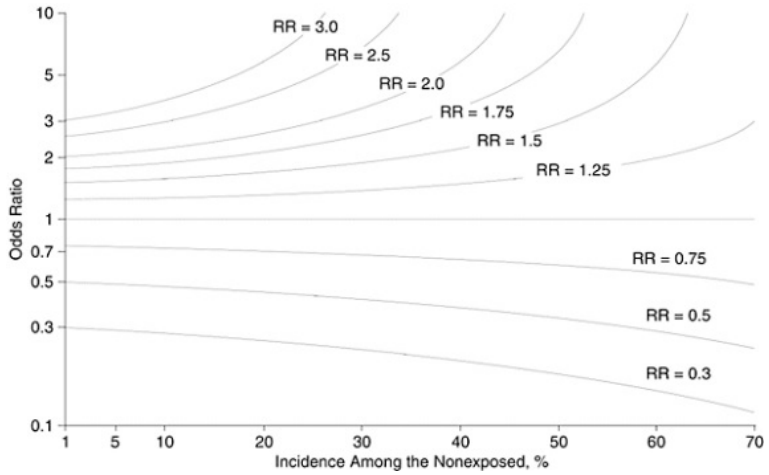


Figura 2.1: Relación entre RR y OR mediante la incidencia entre los no expuestos (Fuente: *Zhang y Yu, 1998*)

La gran ventaja que tienen los OR es que nos permiten el uso de la regresión logística, ya que los parámetros de este modelo pueden interpretarse mediante los términos del *odds ratio*, pero no en los términos del riesgo relativo.

2.3.4 Estimaciones e intervalos de confianza

Consideramos la siguiente tabla que nos facilitará el cálculo de las estimaciones para el OR y el RR:

	D	$\bar{D}$
E	a	b
$\bar{E}$	c	d

donde a es el número de individuos con la enfermedad y expuestos, b el de no enfermos y expuestos, c el de enfermos y no expuestos y d el de no enfermos y no expuestos.

Odds ratio

Asumiendo que los datos vienen de un estudio de cohorte y teniendo en cuenta que la estimación para $\mathrm{P}(D|E)$ será la proporción observada de individuos expuestos que están enfermos, $\hat{p}_1 = a/(a+b)$, y calculando del mismo modo la estimación para $p_0 = \mathrm{P}(D|\bar{E})$, sustituimos estas estimaciones en la fórmula del OR y obtenemos el estimador de máxima verosimilitud (MLE) del *odds ratio*:

$$\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)} = \frac{a \cdot d}{b \cdot c}$$

En cuanto a su intervalo de confianza, nos surge el problema de que la distribución muestral de $\widehat{OR}$ es asimétrica hacia la derecha para muestras pequeñas. Sin embargo, la distribución muestral de $\ln(\widehat{OR})$ es más simétrica y se aproxima mejor mediante una distribución normal cuando el tamaño de la muestra es grande, de modo que aprovechamos esta distribución para calcular los intervalos de confianza del logaritmo, y después será simplemente transformarlo. Sabiendo que:

$$\ln(\widehat{OR}) = \ln\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) - \ln\left(\frac{\hat{p}_0}{1-\hat{p}_0}\right) \sim N(\ln(OR), \mathrm{Var}(\ln(\widehat{OR}))$$

y siguiendo la serie de Taylor de primer orden, llegamos a:

$$\mathrm{Var}(\ln(\widehat{OR})) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

De esta manera, los intervalos de confianza para $\ln(OR)$ y OR, respectivamente, serán:

$$\mathrm{CI}(\ln(OR); 1-\alpha) = \ln(\widehat{OR}) \mp z_{1-\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$\mathrm{CI}(OR; 1-\alpha) = \widehat{OR} \cdot \exp\left(\mp z_{1-\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

Se puede dar el caso de que la muestra con la que trabajamos sea demasiado pequeña, provocando posible sesgo en el estimado $\widehat{OR}$. Para estos casos, se decidió añadir una unidad a los valores del divisor, quedando:

$$\widehat{OR} = \frac{a \cdot d}{(b+1) \cdot (c+1)}$$

Riesgo relativo

Similar a como se ha hecho con la estimación del OR, podemos hallar el MLE del RR, obteniendo:

$$\widehat{RR} = \frac{a/(a+b)}{c/(c+d)}$$

Y de igual forma para su intervalo de confianza, ya que nos surge el mismo problema descrito anteriormente. Por lo que quedaría:

$$\mathrm{CI}(\ln(RR); 1-\alpha) = \ln(\widehat{RR}) \mp z_{1-\alpha/2}\sqrt{\frac{b}{a\cdot(a+b)} + \frac{d}{c\cdot(c+d)}}$$

$$\mathrm{CI}(RR; 1-\alpha) = \widehat{RR}\cdot\exp\left(\mp z_{1-\alpha/2}\sqrt{\frac{b}{a\cdot(a+b)} + \frac{d}{c\cdot(c+d)}}\right)$$

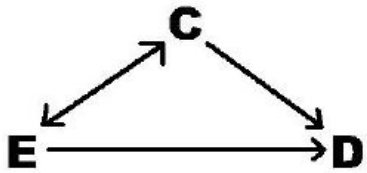
Finalmente, y siguiendo la estructura del OR, cuando nos encontramos con una muestra pequeña, reajustamos el riesgo relativo:

$$\widehat{RR} = \frac{a/(a+b)}{(c+1)/(c+d+1)}$$

2.4 Variables de confusión

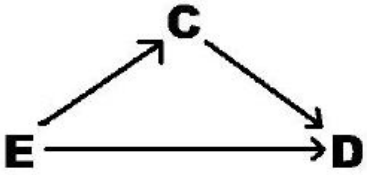
Una de las cosas que no hemos tenido en cuenta en las medidas de asociación comentadas en los puntos anteriores es la presencia de alguna variable de confusión, algo que suele ser muy común y que puede llegar a provocar sesgo en nuestros resultados.

De este modo, una variable de confusión o variable confusora, C, es aquella que está relacionada con E y un factor pronóstico de D y causa una estimación sesgada de la asociación de interés en caso de ser ignorada, pudiendo ser sobrestimada o subestimada. El siguiente gráfico nos ayudará a entender la relación anterior:



La fuerza del sesgo de confusión dependerá de la fuerza de su relación con la exposición y la enfermedad: cuanto mayor es la asociación C-E y la asociación C-D, mayor es el sesgo de confusión.

Conviene remarcar que debemos tener cuidado con la siguiente situación:



en cuyo caso no estaríamos hablando de efecto de confusión, pues si dijéramos aquí que C es variable confusora, es muy posible que no se apreciara bien el efecto de E sobre D en caso de ajustar el análisis por esta variable.

2.5 Regresión logística

Ahora pasamos a hablar acerca de la regresión logística, cuyo modelo es probablemente el más utilizado en el campo de la epidemiología y de la salud pública a la hora de analizar respuestas binarias, y que tendrá gran protagonismo en el siguiente capítulo, donde explicaremos el problema de la separación.

Como se ha dicho anteriormente, es bastante común la existencia de variables confusoras a la hora de cuantificar la asociación entre una enfermedad y una exposición. Un estimador que tiene en cuenta esta existencia es el de Mantel-Haenszel [6], pero este estadístico presenta varias limitaciones, ya que puede no ser adecuado cuando existen más de una posible variable de confusión, si alguna de estas posibles variables es continua o si E es una variable continua. Por este motivo, la regresión logística puede ser una herramienta idónea para analizar el grado de asociación entre E y D.

2.5.1 Definición

Sea $\boldsymbol{X}$ el vector de las variables $X_1, \ldots, X_m$ del modelo, que incluye tanto E como las posibles variables confusoras, y sea Y la variable binaria de interés, que toma valor 1 si se presenta la enfermedad (D) o 0 si hay ausencia ($\bar{D}$). Usamos el modelo de regresión logística, cuyos términos de error siguen una distribución binomial y tiene como función de enlace la *logit*, para modelar la probabilidad condicionada $p = \mathrm{P}(Y = 1|\boldsymbol{X})$ como una función de x, lo que nos permite definir este modelo usando la expresión:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m \tag{2.1}$$

que es equivalente a:

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m)}$$

donde β_0 es la constante y $\beta_1, \ldots, \beta_m$ son los parámetros del modelo, implicando que X_i es un factor de riesgo para Y si $\beta_i > 0$, un factor protector si $\beta_i < 0$ o un término independiente si $\beta_i = 0$.

Cuando tenemos variables categóricas que incluir en nuestro modelo usamos la codificación *dummy*, la cual se compone por una serie de números asignados para indicar la pertenencia en distintos grupos. De esta forma, si uno de los regresores, X_k, es una variable confusora con s niveles, incluiremos s - 1 variables *dummy* en el modelo:

$$X_{k_1} = \begin{cases} 1 & \text{si } X_k = 2 \\ 0 & \text{en otro caso} \end{cases}, \ldots, \; X_{k_{s-1}} = \begin{cases} 1 & \text{si } X_k = s \\ 0 & \text{en otro caso} \end{cases}$$

Se puede coger alguno de los niveles de s como categoría de referencia, pero si X_k es una variable ordinal, es preferible elegir $X_k = 1$ o $X_k = s$ como nivel de referencia siempre que el número de observaciones no sea demasiado pequeño, facilitando así la interpretación del modelo.

2.5.2 Estimación e intervalos de confianza de los parámetros

Podemos estimar los parámetros del modelo de regresión logística usando el *MLE*: dada una muestra de observaciones independientes, $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, la expresión de la función de verosimilitud puede ser expresada como:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}|Y, \boldsymbol{X}) &= \prod_{i=1}^{n} \mathrm{P}\left(Y = y_i|\boldsymbol{x}_i\right) f(\boldsymbol{x}_i) \propto \prod_{i=1}^{n} \mathrm{P}\left(Y = y_i|\boldsymbol{x}_i\right) = \\ &= \prod_{i=1}^{n} \mathrm{P}\left(Y = 1|\boldsymbol{x}_i\right)^{\delta_i} \mathrm{P}\left(Y = 0|\boldsymbol{x}_i\right)^{1-\delta_i} = \\ &= \prod_{i=1}^{n} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)^{\delta_i}}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}\end{aligned}$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)'$ y $\delta_i = 1$ si $Y_i = 1$ y 0 en otro caso. Una vez que obtenemos dicha función, bastaría con maximizar la log-verosimilitud de los datos observados.

Para calcular sus intervalos de confianza utilizamos las propiedades del *MLE* y, bajo la condición de muestras grandes, obtenemos:

$$\hat{\theta}_{ML} \sim N(\theta, \mathrm{Var}(\hat{\theta}_{ML}))$$

donde θ representa cualquiera de los parámetros del modelo $\beta_0, \beta_1, \ldots, \beta_m$.

Por lo tanto, $\mathrm{Var}(\hat{\theta}_{ML})$ es el elemento correspondiente a la diagonal de la inversa de la matriz de información de Fisher y se puede estimar utilizando la observada de esta matriz, por lo que el intervalo de confianza resultante de θ es:

$$\mathrm{CI}(\theta, 1-\theta) = \hat{\theta}_{ML} \mp z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta}_{ML})}$$

2.5.3 Tests de hipótesis

En cuanto a los tests de hipótesis, debemos destacar dos:

- Test de Wald: este test se puede usar para comprobar si una covariable específica X_k está asociada con Y en presencia de las covariables restantes, es decir, se pone a prueba el verdadero valor del parámetro basado en la estimación de la muestra. La hipótesis de este test es $\begin{cases} \mathrm{H}_0 : \beta_k = 0 \\ \mathrm{H}_1 : \beta_k \neq 0 \end{cases}$ y sus dos posibles estadísticos son:

$$\frac{\hat{\beta}_k}{\sqrt{\mathrm{Var}(\hat{\beta}_k)}} \sim^{\mathrm{H}_0} N(0,1); \qquad \frac{\hat{\beta}_k^2}{\mathrm{Var}(\hat{\beta}_k)} \sim^{\mathrm{H}_0} \chi_1^2$$

- Test de razón de verosimilitud: permite verificar la asociación conjunta de varias covariables con Y, donde se contrasta la hipótesis $\begin{cases} \mathrm{H}_0 : \beta_1 = \ldots = \beta_k = 0 \\ \mathrm{H}_1 : \exists j : \beta_j \neq 0 \end{cases}$.
  Para este test se utiliza como estadístico la diferencia de las desvianzas del modelo bajo la hipótesis nula y el modelo completo:

$$\mathrm{Dev}_{\mathrm{H}_0} - \mathrm{Dev}_{full} = -2 \cdot \ln \left(\frac{\mathcal{L}(\hat{\beta}_{0_{\mathrm{H}_0}}, \hat{\beta}_{\mathrm{H}_0} | Y, \boldsymbol{X})}{\mathcal{L}(\hat{\beta}_0, \hat{\beta} | Y, \boldsymbol{X})} \right) \sim^{\mathrm{H}_0} \chi^2_s$$

2.5.4 Interpretación de los parámetros

Para interpretar los parámetros, como dijimos en puntos anteriores, se utiliza el *odds ratio* como medida de asociación. Si tenemos una variable dicotómica, el OR asociado con $X_k = 1$ y ajustado para todas las covariables puede ser expresado mediante:

$$OR_{X_k} = \frac{odds(Y = 1 | X_1, \ldots, X_k = 1, \ldots, X_m)}{odds(Y = 1 | X_1, \ldots X_k = 0, \ldots, X_m)} = \exp(\beta_k)$$

Por tanto, el estimador del OR_{X_k} y el correspondiente intervalo de confianza son:

$$\widehat{OR}_{X_k} = \exp(\hat{\beta}_k)$$

$$\mathrm{CI}(OR_{X_k}; 1 - \alpha) = \widehat{OR}_{X_k} \cdot \exp\left(\mp \, z_{1-\alpha/2} \sqrt{\mathrm{Var}(\hat{\beta}_k)} \right)$$

En el caso de tener una variable continua, el OR asociado que compara dos individuos expuestos que difieren en c unidades es:

$$OR_{X_{k,c}} = \frac{odds(Y = 1 | X_1, \ldots, X_k = x + c, \ldots, X_m)}{odds(Y = 1 | X_1, \ldots, X_k = x, \ldots, X_m)} = \exp(c \cdot \beta_k)$$

Existe también la situación del *odds ratio* asociado con dos covariables, cuyo caso será simplemente multiplicar ambos OR para obtener el conjunto.

Otra interpretación que no debemos olvidar es la de β_0, que está relacionada con la probabilidad de $Y = 1$ en el caso de un individuo con valores cero en todas las covariables, en la que sencillamente tendremos que calcular:

$$p_0 = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

Conviene recordar que esta interpretación que hemos explicado es solo válida para los estudios de cohorte y en los transversales, pero no en los estudios caso-control, debido a que no es posible estimar $\mathrm{P}(Y = 1 | \boldsymbol{X})$. Sin embargo, podemos ajustar un modelo de regresión logística a los datos de un estudio caso-control usando el *MLE* para la estimación de los parámetros. La interpretación de $\boldsymbol{\beta}$ será la misma que en los estudios de cohorte, pero la de la constante del modelo sí que será diferente.

2.5.5 Bondad de ajuste

Una vez que se ha realizado el ajuste del modelo, es conveniente comprobar la bondad de ajuste, y para ello utilizamos el test de Hosmer-Lemeshow [7, 8], en el que la idea general es que, bajo la hipótesis de una especificación correcta del modelo, se espera que el número de eventos predichos por el modelo sea similar a los observados.

Lo que hace este test es ordenar los sujetos de acuerdo al riesgo predicho para la enfermedad, $\hat{p} = \mathrm{P}(Y = 1|\boldsymbol{X})$ en el caso de estudios de cohorte, y los divide en grupos de 5 a 10 aproximadamente del mismo tamaño. Entonces con cada uno de estos g grupos, el número de eventos observados, O_k con $k = 1, \ldots, g$, es comparado con el número esperado, E_k. El estadístico que se usa para este caso es:

$$\chi^2_{HL} = \sum_{k=1}^{g} \frac{(O_k - E_k)^2}{V_k} \sim^{\mathrm{H}_0} \chi^2_{g-2}$$

La gran desventaja que presenta este test es que depende del número de grupos y de cómo los sujetos estén asignados a estos grupos en caso de empate, además de que no se trata de un test muy robusto.

Capítulo 3

El problema de la separación

En este capítulo explicaremos en qué consiste el problema de la separación, un fenómeno que aparece con más frecuencia de lo que pensamos y que está directamente relacionado con la regresión logística cuando los resultados son dicotómicos o categóricos.

A lo largo de dicho capítulo veremos, junto con la ayuda de algunos pequeños ejemplos, las consecuencias que abarca este problema, tanto en la teoría como en la práctica, y las posibles soluciones que se han desarrollado para subsanarlo, llegando al método que propuso Firth, el cual es hasta día de hoy el método que más se utiliza en la corrección de este fenómeno.

3.1 Motivación

Para motivar este capítulo, lo primero que vamos a ver es un pequeño ejemplo ficticio donde se produce separación. Por tanto, consideramos 20 individuos, de los cuales la mitad tienen la enfermedad de estudio (D) y la otra mitad no tienen dicha enfermedad ($\bar{D}$), mientras que 5 están expuestos (E) y los otros 15 no ($\bar{E}$), de tal manera que quedan distribuidos de la siguiente manera:

	D	$\bar{D}$	Total
E	5	0	5
$\bar{E}$	5	10	15
Total	10	10	20

Procedemos a realizar el cálculo manual del estimador del *odds ratio* y vemos que tiende a infinito:

$$\widehat{OR} = \frac{5 \cdot 10}{0 \cdot 5} = \frac{50}{0} \longrightarrow \infty$$

Ahora pasamos a ver el resultado que obtenemos al realizar un análisis estándar con R para poder compararlo con lo que nos ha dado anteriormente. Para ello, vamos a utilizar la función *glm* [27]:

```
> modelo_glm <- glm(D~E, datos_ejemplo, family = 'binomial')
> summary(modelo_glm)

Call:
glm(formula = D ~ E, family = "binomial", data = datos_ejemplo)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.17741  -0.29441  -0.00008   0.29429   1.17741

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -19.57    3400.72  -0.006    0.995
E1             19.57    3400.72   0.006    0.995

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.493  on 19  degrees of freedom
Residual deviance: 13.863  on 18  degrees of freedom
AIC: 17.863

Number of Fisher Scoring iterations: 18
```

Una vez que tenemos ambos resultados, podemos ver que hay algún problema, ya que existen dos soluciones diferentes: por una parte tenemos que $\widehat{OR}$ diverge a infinito y, por otra parte, que esta misma estimación es $e^{19,57} = 315604372$, un valor bastante alto pero que no es comparable con infinito. También cabe destacar el error estándar tan grande que nos da el programa, cerca de 3400, algo que es inaceptable. Todo esto son consecuencias que vienen derivadas del problema de la separación.

3.2 La función de verosimilitud

Antes de que pasemos a explicar el problema de la separación, es conveniente que introduzcamos algunas notaciones importantes sobre la función de verosimilitud y su estimador, así como mencionar la verosimilitud penalizada [9, 10], ya que nos ayudará a entender mejor algunos conceptos que veremos más adelante.

3.2.1 Estimador de máxima verosimilitud

El método de máxima verosimilitud (ML) para la estimación de parámetros en modelos de regresión es ampliamente utilizado en epidemiología pero son muchos los

epidemiólogos que reciben poca o ninguna educación en los fundamentos conceptuales del enfoque. Al igual que con todos los métodos estadísticos inferenciales, el de máxima verosimilitud se basa en un modelo supuesto y no puede dar cuenta de las fuentes de sesgo que no están controladas por el modelo o el diseño del estudio, por lo que hay que tener cuidado con el modelo que se propone, ya que si dicho modelo no es adecuado, el método tampoco lo será. Sin embargo, y a pesar de sus pequeños inconvenientes, este método es muy popular, porque es computacionalmente directo e intuitivo, y porque los estimadores de máxima verosimilitud (*MLE*) tienen propiedades deseables en muestras grandes en el caso en el que el modelo se haya especificado correctamente.

De esta forma, sea $\boldsymbol{X}$ un vector de las variables $X_1, \ldots, X_m$, sea Y una variable binaria de interés y sea $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)'$ el vector de los parámetros del modelo (2.1), podemos calcular los *MLE*, que en caso de que existan serán finitos y únicos, realizando los siguientes pasos:

1. Escribimos la función de verosimilitud: $\mathcal{L}(\boldsymbol{\beta}|\, Y, \boldsymbol{X})$
2. La convertimos en la función de log-verosimilitud: $l(\boldsymbol{\beta}|\, Y, \boldsymbol{X}) = \ln \mathcal{L}(\boldsymbol{\beta}|\, Y, \boldsymbol{X})$
3. Obtenemos el vector *score*: $U(\beta_k) \equiv \dfrac{\partial}{\partial \beta_k}\, l(\boldsymbol{\beta}|\, Y, \boldsymbol{X})$
4. Igualamos a cero este vector para obtener los posibles estimadores: $U(\beta_k) = 0$
5. Comprobamos que realmente son máximos: $\dfrac{\partial^2}{\partial \beta_k^2}\, l(\boldsymbol{\beta}|\, Y, \boldsymbol{X}) < 0$

Este método puede llegar a tener algunas dificultades en cuanto a estabilidad, cuya solución puede ser la verosimilitud penalizada [11].

3.2.2 Verosimilitud penalizada

La verosimilitud penalizada es un método para eludir los problemas en la estabilidad de las estimaciones de los parámetros que surgen cuando la verosimilitud es relativamente plana, lo que dificulta la determinación de los *MLE* por medio del enfoque estándar.

La estimación de verosimilitud penalizada, también conocida como estimación *shrinkage*, se caracteriza porque tiene en cuenta la complejidad del modelo cuando se estiman los parámetros de diferentes modelos. Básicamente, en lugar de hacer una simple estimación de máxima verosimilitud, maximiza la log-verosimilitud menos un término de penalización, que depende del modelo y generalmente aumenta con el número de parámetros. Desde otra perspectiva, se puede ver como un método para introducir cierto grado tolerable de sesgo a cambio de la reducción en la variabilidad de las estimaciones de los parámetros, lo que dará un mejor ajuste a los datos y, por

consiguiente, una mayor verosimilitud. Cabe mencionar que la penalización se puede aplicar a cualquier método de estimación.

3.3 La separación en regresión logística

Como dijimos en el capítulo anterior, la regresión logística es una de las técnicas estadísticas más aplicadas cuando se busca explicar el comportamiento probabilístico de algún fenómeno, lo que ha hecho que se convierta en una herramienta de uso permanente entre investigadores de la salud; y un problema que aparece con frecuencia en estos modelos es el de la separación de los datos [12, 13, 14, 15].

3.3.1 Definición

El problema de la separación, o verosimilitud monótona, es un fenómeno que se encuentra en los modelos de regresión con un resultado dicotómico o categórico y donde se muestran los grupos de éxito separado de los fracasos, es decir, cuando las respuestas y no respuestas están perfectamente separadas por algún factor de riesgo o una combinación lineal de factores de riesgo.

La separación no es un tema minúsculo, ya que en la teoría, aunque la función de verosimilitud converja, puede producir estimaciones infinitas o erróneas para algunos coeficientes. En la práctica el problema incrementa, pues puede pasar desapercibida o mal manejada debido a los límites del software para reconocer y manejar este fenómeno, por lo que muchas veces los investigadores no son conscientes de la existencia de este hecho y presentan resultados que pueden no ser adecuados a los estimadores reales.

Cabe decir que el problema de la separación en regresión logística se suele presentar principalmente en estudios con muestras pequeñas y datos dispersos con varios factores de riesgo altamente predictivos y no balanceados. También es frecuente encontrarlo en estudios con presencia de un resultado raro, exposiciones raras, covariables altamente correlacionadas o covariables con fuertes efectos.

Es por esto que la probabilidad de separación dependerá, como vemos en la Tabla 3.1 con un ejemplo, del tamaño de la muestra, del número de factores dicotómicos, de la magnitud de los OR asociados a estos factores y del grado de balanceo de los grupos: cuanto menor es la muestra, más factores de riesgo hay en el modelo, mayor es la magnitud de los OR y menor es el balanceo de la base de datos, mayor será la probabilidad de separación en los datos.

		Grado de balanceo de los factores de riesgo							
		1:1				1:4			
		OR				*OR*			
Tamaño muestral	Número de factores de riesgo	1	2	4	16	1	2	4	16
30	3	0	3	10	53	17	25	43	74
	5	2	7	24	75	30	41	58	85
	10	12	38	78	98	56	71	86	98
50	3	0	0	1	18	2	5	15	46
	5	0	0	2	32	6	9	22	53
	10	0	1	20	78	10	19	36	74

Tabla 3.1: Probabilidad de separación (en porcentaje) en regresión logística con factores de riesgo dicotómicos. Cada entrada está basada en 1000 muestras (Fuente: *Heinze y Schemper, 2002*)

3.3.2 Tipos de separación

En cuanto a los tipos de separación que podemos encontrar, existe cierta controversia de como clasificar los datos que se han recogido para un estudio, pero algunos autores, como *Albert y Anderson (1984)* [16], defienden que solo existen tres posibles categorías mutuamente exclusivas y exhaustivas: dos tipos de separación (separación completa y cuasicompleta) y los datos sobrepuestos (*Overlap*). Pasamos a explicar cada categoría:

- **Separación completa**

  Hablamos de separación completa (Figura 3.1) cuando, como su propio nombre indica, se presenta una división completa de los dos grupos de puntos asociados a los valores que toma la variable respuesta (adoptando una codificación general de 0 y 1): uno de los grupos será donde no ocurre el evento de interés y el otro grupo donde ocurre dicho evento. En el caso de una sola variable explicativa x, la separación se presenta cuando ocurren todos los fracasos en la primera parte del rango de la variable x (R1) y todos los éxitos en la segunda parte de este rango (R2), o viceversa, sin dar lugar a una sobreposición de ambos rangos, o mezcla de éxitos y fracasos. Sin embargo existe un tercer rango de x, donde no hay realizaciones de la variable Y, y es conocida como la región de separación, ya que separa totalmente los éxitos de los fracasos (RS).

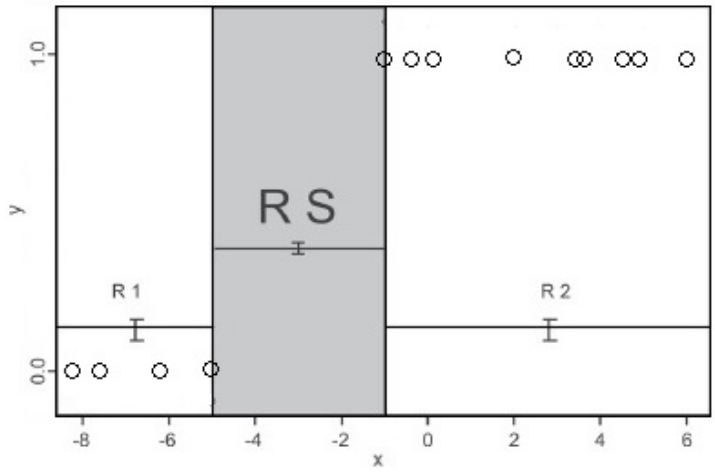


Figura 3.1: Ejemplo de datos con separación completa

- **Separación cuasicompleta**

  Este concepto fue desarrollado más tarde que el descrito anteriormente [17], y decimos que existe separación cuasicompleta (Figura 3.2) cuando es posible definir un plano que pasa por la región de separación con éxitos a un lado o sobre este y fracasos al otro o sobre este, sin presentarse convergencia de los estimadores de máxima verosimilitud.

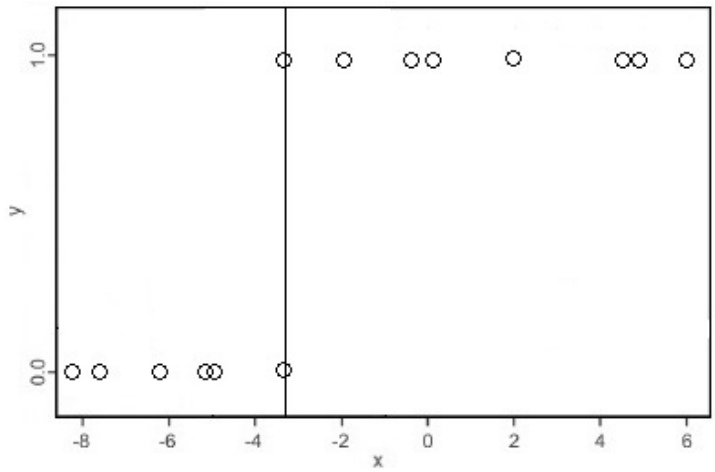


Figura 3.2: Ejemplo de datos con separación cuasicompleta

- ***Overlap***

  En este caso, no existe ninguna recta, plano o región de separación entre los datos, por lo que se encuentran mezclados o sobrepuestos en todo el rango de valores de x (Figura 3.3). Los datos que presentan *overlap* serán aquellos que no den problemas a la hora de calcular los *MLE*.

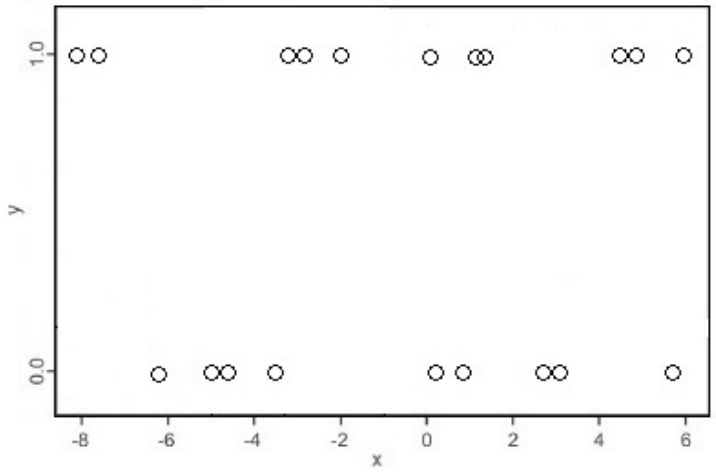


Figura 3.3: Ejemplo de datos con *Overlap*

3.3.3 Ejemplo

Para facilitar la comprensión de los conceptos anteriores, vamos a ver un ejemplo real donde hay separación [14]. En este, utilizamos un conjunto de datos sobre 907 jóvenes de la ciudad de Medellín (Colombia), tomados en el año 2004, con edades comprendidas de 5,1 a 19,5 años. A las jóvenes se les preguntó si ya habían presentado o no menarquía (primer episodio menstrual de la mujer), obteniendo los datos que se ven en la Tabla 3.2.

	Rangos de edades					
Menarquía	5,07 - 7	7 - 10,3	10,3 - 14,4	14,4 - 19,5	Cantidad de jóvenes	%
No	132	411	0	0	543	59,87
Sí	0	0	0	364	364	40,13

Tabla 3.2: Datos de la edad de menarquía (Fuente: *Correa y Valencia, 2011*)

En la tabla anterior observamos que hasta los 10,3 años ninguna joven había presentado menarquía, entre las edades 10,3 y 14,4 no hay datos, y después de los 14,5 años, todas habían presentado ya la menarquía. Por consiguiente, apreciamos claramente que los datos presentan separación completa y que la región de separación va de 10,3 a 14,5 años.

Se procede a analizar estos datos a través del programa R, y nos da los resultados que se muestran en la Tabla 3.3.

	Coeficientes			
	Estimación	Error estándar z	valor	Pr(> |z|)
(intercept)	-130,87	39296,15	-0,003	0,997
edadcal	10,56	3177,83	0,003	0,997

Tabla 3.3: Resultados aproximados para los parámetros a través de R (Fuente: *Correa y Valencia, 2011*)

Podemos observar que el programa nos devuelve la estimación de la constante del modelo y de β_1, que es el parámetro correspondiente a la edad en la que se presenta la menarquía. Estas estimaciones son datos aproximados, pero vemos que carecen de sentido ya que sus errores estándar son exageradamente grandes, lo cual se debe a la existencia de separación.

3.4 Posibles soluciones al problema de la separación

Lo primero que recomendamos hacer es aplicar el método *ML* para detectar posibles problemas, y en caso de que nos encontremos separación en un conjunto de datos, debemos aclarar si el problema se puede eliminar mediante una revisión sensata de dichos datos, ya que ejemplos típicos de estrategias que dan lugar a la separación son la categorización de variables continuas o el uso de muchas categorías para variables nominales. A continuación, vamos a suponer que la separación no se puede eliminar revisando los datos.

Por lo tanto, si se detecta separación causada por algún factor de riesgo determinado, entonces existen varias alternativas a proceder:

1. Sacar el factor de riesgo del modelo.
2. Cambiar el tipo de modelo estimado.
3. Hacer un ajuste *ad hoc* para la manipulación de los datos [15].
4. Realizar un análisis estándar y considerar el coeficiente del factor de riesgo problemático como un valor alto en lugar de infinito.

Estas opciones son inmediatas y muy fáciles de aplicar, pero presentan inconvenientes que pueden incluso llegar a agravar más el problema:

1. Esta opción no es recomendada porque no proporciona información sobre el efecto de este factor de riesgo importante y tampoco nos permite ajustar los efectos de los otros factores de riesgo para dicho factor.

2. Expresar los factores de riesgo en términos de $\ln(OR)$ no solo es común en el análisis de respuestas binarias, sino que también es útil y fácil de interpretar. El problema es que no en todos los tipos de modelos los coeficientes son fácilmente interpretables.

3. Los ajustes *ad hoc* que proceden a un análisis estándar pueden producir estimaciones finitas añadiendo respuestas artificiales y no respuestas artificiales para cada grupo de los diferentes factores de riesgo en la base de datos, y luego realizar un análisis estándar en el conjunto de datos aumentado. El inconveniente de este método es que no se conoce bien las propiedades de este procedimiento, ya que todavía no se ha llegado a desarrollar al 100 %.

4. Cambiar el coeficiente del factor de riesgo problemático que resulta infinito por un valor alto implica una inflación extrema de la varianza de este coeficiente. Esto conduce a un test de Wald insignificante que puede no ser plausible debido a un efecto muy fuerte. Se puede calcular el intervalo de confianza del estimador por otros métodos, pero el hecho de elegir arbitrariamente un valor alto para el estimador sigue siendo insatisfactorio.

Otras alternativas más complejas, pero más próximas a la corrección del problema de la separación, son el uso de un modelo de regresión logística oculto y el uso de la regresión logística exacta.

3.4.1 Modelo de regresión logística oculto

Este modelo fue propuesto por *Rousseeuw y Christmann (2003)* [18] como alternativa al modelo de regresión cuando se produce el efecto de separación. Se trata de un modelo más general bajo el cual la respuesta observada está fuertemente relacionada pero no es igual a la respuesta verdadera no observable, es decir, las respuestas no observadas se consideran como latentes.

Este método consiste en calcular unas pseudo-observaciones $\tilde{y}_i$, para posteriormente ajustar el modelo de regresión logística sustituyendo las observaciones iniciales por dichas pseudo-observaciones. El *MLE* resultante siempre existe y es único, pero el problema que presenta es que, como hablamos de un modelo generalizado, muchas veces estos estimadores difieren bastante de la realidad.

3.4.2 Regresión logística exacta

Existen muchos documentos que hablan sobre la regresión logística exacta, y son bastantes los que coinciden en que es la alternativa lógica a la máxima verosimilitud [19, 20]. Estos modelos se utilizan cuando el tamaño de la muestra es demasiado pequeño para analizarlos con los de la regresión logística regular y/o cuando algunas

de las celdas formadas por el resultado y la variable predictora categórica no tienen observaciones, abarcando así el problema de la separación.

La regresión logística exacta fue desarrolla para obtener tests de los coeficientes de regresión para los cuales se garantiza que las probabilidades de tipo I no excedan los niveles nominales. La inversión de estos tests exactos produce intervalos de confianza con propiedades análogas.

La ventaja que nos concede este método es que puede proporcionar estimaciones puntuales finitas a través de la mediana insesgada, sustituyendo al *MLE* no adecuado. Pero desafortunadamente, estas estimaciones pueden comportarse inesperadamente con datos extremadamente dispersos. También presenta el inconveniente de que, aunque puede llegar a mejorar a la regresión logística en el caso de un predictor, puede dar resultados erróneos cuando hay más de un parámetro a la vez o cuando las variables explicativas son continuas.

Otro comentario negativo a tener en cuenta dentro de la regresión logística exacta es que computacionalmente, con los programas estadísticos existentes hasta la fecha, requiere mucha capacidad de memoria y tiempo de cálculo, lo que lo hace inviable muchas veces [21].

3.5 La función *score* modificada de Firth

Acabamos de ver que ninguna de las posibles soluciones propuestas en el punto anterior llega a corregir del todo el problema de la separación, ya que todas presentan serias limitaciones que ponen en duda su uso en la práctica.

Las modificaciones de la verosimilitud pueden proporcionar mejores soluciones. Si la función de log-verosimilitud a maximizar se modifica ligeramente añadiendo un término de penalización adecuado, se pueden evitar estimaciones infinitas. Varias de estas penalizaciones fueron sugeridas y motivadas por distintos objetivos [22]: algunas reducen el sesgo o el error cuadrático medio de las estimaciones, y otras están motivadas desde una perspectiva bayesiana, incorporando información de que los coeficientes infinitos son incorrectos.

De esta manera, surgió el método de reducción de sesgo [12, 23], que consiste básicamente en la modificación de la función *score* de modo que las raíces de las ecuaciones *score* modificadas resultantes son estimadores insesgado de primer orden. David Firth fue el primero en dar una base formal a este método, cuya idea inicial era reducir el sesgo de los *MLE* introduciendo un pequeño sesgo en la función *score*. Demostró que para la familia exponencial de parametrización canónica, el método de verosimilitud penalizada se reduce a la priori invariante de Jeffreys.

Las ventajas de este método son:

- Es independiente del *MLE*, por lo tanto no depende de su finitud, lo que hace que sus estimadores sean siempre finitos.

- Los nuevos estimadores obtenidos tienen el término de primer orden más pequeño, o incluso cero, en la expansión asintótica de su sesgo.
- Estos estimadores, de sesgo reducido, al ser obtenidos por estimación *ML* tienen todas las propiedades asintóticas deseadas: normalidad asintótica, suficiencia asintótica, insesgamiento y eficiencia.

Este enfoque ganó popularidad debido a las propiedades superiores del estimador de sesgo reducido sobre el tradicional *MLE*, generalmente en modelos para respuestas categóricas, y particularmente en regresión logística.

3.5.1 Descripción del método

Como ya dijimos, los *MLE* de los parámetros de regresión se obtienen como soluciones a las ecuaciones *score*, $U(\beta_k)$. Con el fin de reducir el pequeño sesgo muestral de estos estimadores, *Firth (1993)* sugirió basar la estimación en ecuaciones *score* modificadas:

$$U(\beta_k)^* \equiv U(\beta_k) + \frac{1}{2} \cdot trace \left[I(\beta)^{-1} \left(\frac{\partial I(\beta)}{\partial \beta_k} \right) \right] = 0$$

con $k = 1, \ldots, m$ y donde $I(\beta)^{-1}$ es la inversa de la matriz de información de Fisher evaluada en β.

En la Figura 3.4 podemos ver visualmente la relación existente entre la función *score* y la función *score* modificada.

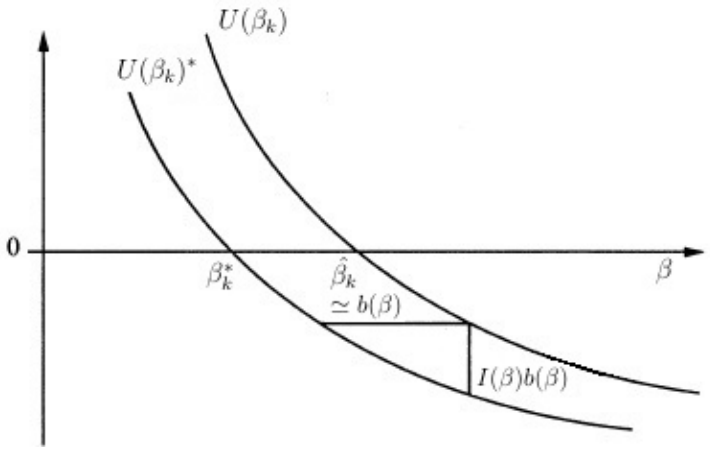


Figura 3.4: Relación entre la función *score* y la función *score* modificada (Fuente: *Firth, 1993*)

Cabe decir que la función *score* modificada está relacionada tanto con la función de verosimilitud penalizada, $L(\beta^*) = L(\beta)|I(\beta)|^{1/2}$, como con la función de

log-verosimilitud penalizada, $\ln L(\beta^*) = \ln L(\beta) + \frac{1}{2} \cdot |I(\beta)|$, siendo $|I(\beta)|^{1/2}$ la priori invariante de Jeffreys para este problema. Su influencia es asintóticamente insignificante y al usar esta modificación, Firth mostró que se elimina el sesgo $b(\beta)$ de los *MLE*.

Si la idea general de Firth es aplicar esta modificación a un modelo logístico:

$$p_i = \mathrm{P}(y_i = 1|\boldsymbol{x}_i, \beta) = \left[1 + \exp\left(-\sum_{r=1}^{n} \boldsymbol{x}_{ik}\beta_k\right)\right]^{-1}$$

entonces basta con reemplazar la ecuación *score* $U(\beta_k) = \sum_{i=1}^{n}(y_i - p_i)\boldsymbol{x}_{ik} = 0$ por la ecuación *score* modificada:

$$U(\beta_k)^* = \sum_{i=1}^{n}\left[y_i - p_i + h_i\left(\frac{1}{2} - p_i\right)\right]\boldsymbol{x}_{ik} = 0$$

donde los h_i son los i-ésimos elementos de la diagonal de la *hat matrix*:

$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$$

con $W = diag\,[p_i(1 - p_i)]$. Ahora las nuevas estimaciones se pueden obtener iterativamente de la manera habitual hasta que se obtenga la convergencia:

$$\beta^{(s-1)} = \beta^{(s)} + I^{-1}(\beta^{(s)})U(\beta^{(s)})^*$$

donde el superíndice (s) se refiere a la iteración s-ésima.

Alternativamente al método que hemos presentado, los estimadores calculados por el método de Firth se pueden obtener dividiendo cada observación original i en dos nuevas observaciones que tienen valores respuesta y_i y $1 - y_i$, con pesos actualizados iterativamente $1 + h_i/2$ y $h_i/2$, respectivamente. Esta división garantiza que haya estimaciones finitas y, por lo tanto, elimina completamente el problema de separación. Además, estos nuevos estimadores generalmente son más pequeños en valor absoluto que los *MLE*, por lo que sus errores estándar se reducirán también.

3.5.2 Un estudio empírico

Se llevó a cabo un estudio de simulación, realizado por *Heinze y Schemper (2002)*, donde podemos explorar y comparar el rendimiento empírico de los tres ajustes más relevantes que hemos visto en este capítulo: la máxima verosimilitud estándar (*ML*), el propuesto por Firth (*FL*) y el de la regresión logística exacta (*XL*). En la Tabla 3.4 podemos observar los resultados obtenidos.

			Grado de balanceo de los factores de riesgo							
			1:1				1:4			
			OR				*OR*			
			1	2	4	16	1	2	4	16
			Valor del parámetro = $\ln(OR)$							
Tamaño muestral	Número de factores de riesgo	Método	0	0,69	1,39	2,77	0	0,69	1,39	2,77
30	3	*ML*	-4	32	102	566	-7	88	186	424
		FL	-3	1	1	-6	-2	-1	-5	-19
		XL	-3	4	6	-35	-2	-2	-10	-42
	10	*ML*	-27	574	1118	1168	-8	326	897	1292
		FL	0	3	-23	-130	2	8	-6	-89
100	3	*ML*	0	4	10	34	1	5	9	58
		FL	0	1	2	2	1	-2	-3	-1
	10	*ML*	1	11	34	429	1	15	32	233
		FL	1	0	2	8	1	3	4	5

Tabla 3.4: Sesgo promedio $\times$ 100 de las estimaciones de los parámetros en la regresión logística. Cada entrada se basa en 1000 muestras (Fuente: *Heinze y Schemper, 2002*)

Lo primero que nos llama la atención de los datos es que el método de regresión logística exacta no está disponible para el tamaño muestral 100 ni para $n = 30$ cuando el número de factores de riesgo es 10. Esto se debe a las distribuciones condicionales degeneradas y a que, como dijimos anteriormente, los requisitos de memoria y tiempo de cálculo para este método son excesivos.

También vemos claramente que el sesgo del estimador, que depende del valor del parámetro, varía en función del tamaño de la muestra, del número de factores de riesgo, de la magnitud del *odds ratio* y del balanceo de los datos. En concreto, como podemos ver más explícitamente en la Tabla 3.5, hay más sesgo cuanto menor tamaño muestral haya, mayor número de factores, mayor magnitud de los *OR*, menor balanceo de los datos y mayor valor del parámetro. Por lo tanto, observamos que el sesgo del estimador, en valor absoluto, con *FL* es pequeño, que es más grande con *XL* y que aún es más grande con *ML*, donde teóricamente debería ser infinito si ocurre la separación.

Tamaño muestral	Número de factores de riesgo	*OR*	Grado de balanceo	Valor del parámetro	Sesgo
+ grande	− factores	− magnitud	+ balanceo	+ pequeño	− sesgo
+ pequeño	+ factores	+ magnitud	− balanceo	+ grande	+ sesgo

Tabla 3.5: Relación del sesgo según el tamaño muestral, el número de factores, el *odds ratio*, el balanceo y el valor del parámetro

Resumiendo, el estudio nos confirma el uso seguro de *FL* en general y su clara superioridad sobre el *ML* y el *XL*, particularmente en situaciones de valores de parámetros y/o factores de riesgo no balanceados.

Capítulo 4

Ejemplo de separación con datos reales

En el presente capítulo realizaremos un análisis estadístico con R sobre un estudio de cáncer de endometrio, donde aplicaremos el método de máxima verosimilitud frecuente y el método que hemos presentado en el capítulo anterior expuesto por Firth, para posteriormente compararlos. Pero antes de esto explicaremos, en forma de pequeña guía, todo lo que necesitamos saber acerca de este tipo de cáncer [24, 25, 26].

4.1 Cáncer de endometrio

4.1.1 Definición

El endometrio es el revestimiento interior del útero (Figura 4.1), un órgano hueco y muscular de la pelvis de la mujer que tiene aproximadamente el tamaño y la forma de una pera, y donde crece y se desarrolla el feto.

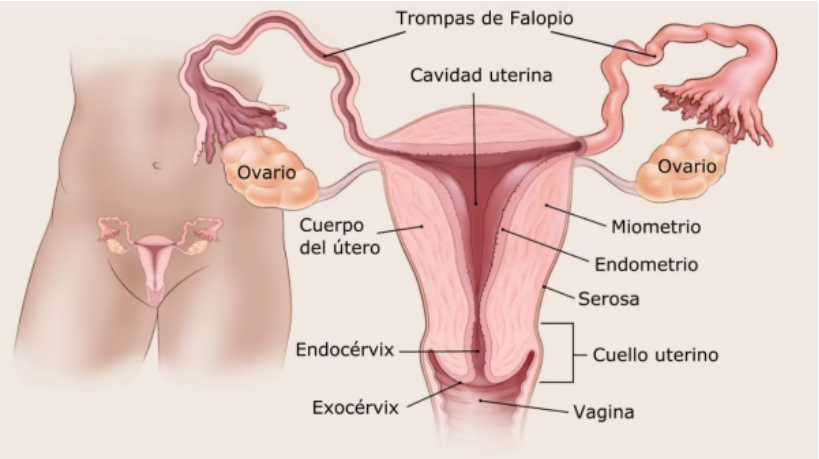


Figura 4.1: Anatomía del aparato reproductor femenino (Fuente: ACS)

Por lo que definimos el cáncer de endometrio, que es el tipo más común de cáncer uterino, como una enfermedad que se origina cuando células malignas (cancerosas) comienzan a crecer en forma descontrolada en este revestimiento.

4.1.2 Factores de riesgo

Todavía no se saben las causas exactas de esta enfermedad, pues las células de casi cualquier parte del cuerpo pueden convertirse en cáncer y extenderse a otras áreas del cuerpo, pero sí que se conocen algunos factores de riesgo, donde uno de los principales es el desequilibrio hormonal, ya que la mayoría de las células cancerosas endometriales contienen receptores de hormonas de estrógeno y/o progesterona en sus superficies y, de alguna manera, la interacción de estos receptores con sus hormonas conduce a un mayor crecimiento del endometrio. Esto podría ser el inicio de cáncer.

Para entender mejor cómo se producen las hormonas de estrógeno y progesterona, es apropiado que expliquemos brevemente el ciclo menstrual de una mujer. En la etapa inicial, antes de la ovulación, los ovarios producen estrógenos, los cuales causan que el endometrio se haga más grueso para que pueda nutrir a un embrión en caso de ocurrir un embarazo. Si no surge el embarazo, estas hormonas se producen en menores cantidades, dando lugar a que se produzcan las progesteronas, después de la ovulación. Esto prepara a la capa interior del revestimiento para ser eliminada y desechada del útero al final del ciclo, que es lo que llamamos regla o período. Este ciclo se repite hasta que la mujer pasa por la menopausia, momento en el que pasa también a tener mayor desequilibrio hormonal.

Por esta razón, el cáncer de endometrio se desarrolla tanto en mujeres pre-menopáusicas (25 %) como en mujeres post-menopáusicas (75 %), ocurriendo la gran mayoría de los casos entre los 50 y los 69 años.

Otros factores a tener en cuenta en el riesgo de padecer este cáncer pueden ser:

- Factores que afectan los niveles hormonales, tales como tomar estrógeno después de la menopausia, píldoras anticonceptivas o tamoxifeno.
- Índice de pulsación de la arteria uterina.
- Altura del endometrio.
- Diabetes, anemia u obesidad.
- Edad a la que se presenta la primera menstruación y la menopausia.
- Presencia o ausencia de neovasculización.
- Tener parientes cercanos con cáncer de endometrio o cáncer colorrectal.
- Haber sido diagnosticada con cáncer de seno o de ovario en el pasado.

- Haber sido diagnosticada con hiperplasia endometrial en el pasado.
- Tratamiento con radioterapia a la pelvis para tratar otro cáncer.

Aunque cabe decir que se puede dar el caso de mujeres con cáncer de endometrio que no presenten ninguno de los factores de riesgo conocidos hasta ahora. Incluso si una mujer con esta enfermedad presenta uno o más factores de riesgo, no hay forma de saber cuáles de estos factores han sido responsables de su cáncer.

4.1.3 Clasificación

El grado de un cáncer de endometrio se basa en la cantidad de glándulas que forma el cáncer que lucen similares a las encontradas en el endometrio normal y saludable. De manera que podemos clasificarlos en tres grados:

- Grado I: los tumores tienen 95 % o más de tejido canceroso que forma glándulas.
- Grado II: los tumores tienen entre 50 % y 94 % de tejido canceroso que forma glándulas.
- Grado III: los tumores tienen menos de la mitad de tejido canceroso que forma glándulas y tienden a ser agresivos y a tener peor pronóstico.

Existe también una clasificación según la propagación del cáncer a otras partes del cuerpo:

- Tipo 1: abarca los tumores de grado I y II. Por lo general, estos cánceres no son muy agresivos y no se propagan rápidamente a otros tejidos (Figura 4.2).

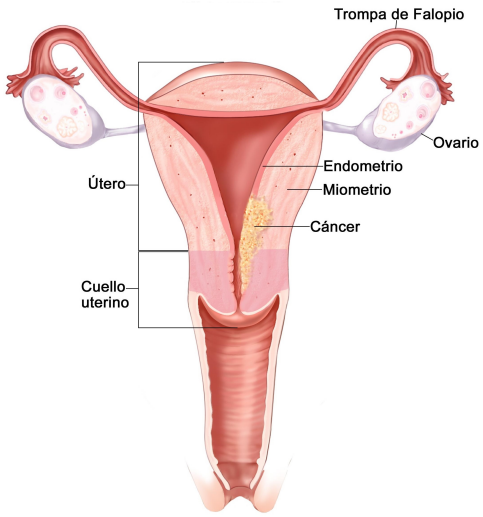


Figura 4.2: Cáncer de endometrio de tipo 1 (Fuente: NIH)

- Tipo 2: son los correspondientes a los de grado III, y tienen una probabilidad mayor de crecer y propagarse fuera del útero, con un pronóstico más desfavorable (Figura 4.3).

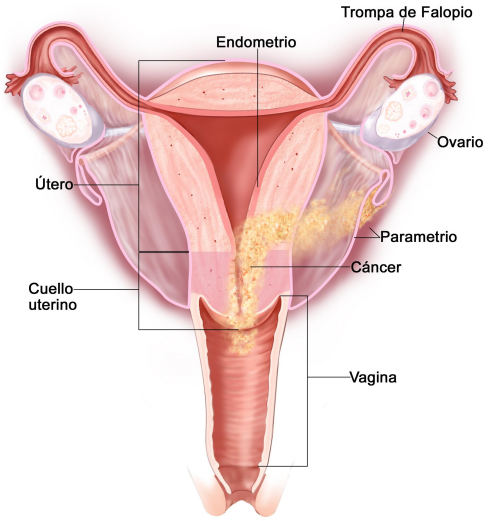


Figura 4.3: Cáncer de endometrio de tipo 2 (Fuente: NIH)

4.1.4 Síntomas y prevención

Son varios los síntomas, en caso de que se manifiesten, que puede presentar una paciente que tenga cáncer endometrial, pero tres son los más frecuentes:

1. Sangrado anormal de la vagina, incluyendo sangrado entre períodos normales o manchado/sangrado después de la menopausia.

2. Episodios de sangrado vaginal frecuentes, fuertes o extremadamente prolongados después de los 40 años.

3. Dolor abdominal bajo o calambres pélvicos.

En cuanto a la prevención, existen algunos factores de riesgo para este tipo de cáncer que no se pueden evitar, como la edad de la persona o heredar ciertos genes, pero hay muchos otros factores que minimizan la probabilidad de enfermar, o disminuyen la gravedad en su caso, como son hacer ejercicio con regularidad, consumir alimentos saludables o realizar controles médicos períodicamente.

4.2 Análisis estadístico con R

En este punto procedemos a realizar un análisis estadístico sobre el cáncer de endometrio pero, antes de nada, debemos denotar que el conjunto de datos con el que

vamos a trabajar fue elaborado por la Profesora y Doctora Ella Asseryanis, de la Universidad de Viena, y se encuentra disponible dentro del paquete *brglm2* de R [28].

Este estudio, que se compone por 79 mujeres diagnosticadas principalmente con cáncer endometrial, tiene como objetivo explicar el grado de histología del endometrio por medio de unos supuestos factores de riesgo.

La histología (HG) se clasifica según el grado en el que se encontraba el cáncer en el momento en el que se examinó a la paciente, de tal manera que:

$$HG = \begin{cases} 0 & \text{si es de grado I–II} \\ 1 & \text{si es de grado III} \end{cases}$$

Mientras que los posibles factores de riesgo que se tuvieron en cuenta fueron tres:

- Un factor dicotómico: la neovasculización (NV), que se codificó como 0 si estaba ausente y 1 en caso de presencia.
- Dos factores continuos: el índice de pulsación de la arteria uterina (PI) y la altura del endometrio en cm (EH).

Pasamos pues a analizar los datos, y lo primero que vamos a hacer es ver cómo están estructuradas las dos variables dicotómicas (Tabla 4.1) y elaborar un resumen descriptivo para las dos variables continuas (Tabla 4.2).

	HG		*NV*	
	Grado I–II	Grado III	Ausente	Presente
Frecuencia	49	30	66	13
Porcentaje (%)	62,03	37,97	83,54	16,46

Tabla 4.1: Frecuencias de HG y NV

	PI	*EH*
Mínimo	0,00	0,27
1er Cuartil	11,00	1,18
Mediana	16,00	1,64
Media	17,38	1,66
3er Cuartil	21,00	2,02
Máximo	49,00	3,61

Tabla 4.2: Resumen descriptivo de PI y EH

Ahora nos centramos en ver si hay un posible problema de separación entre cada uno de los tres factores con respecto a la variable respuesta *HG*. Para ello, un diagrama de cajas nos facilitará visualmente si existe dicho problema para los dos factores continuos (Figura 4.4), mientras que para la variable *NV*, bastará con hacer una tabla de contingencia (Tabla 4.3).

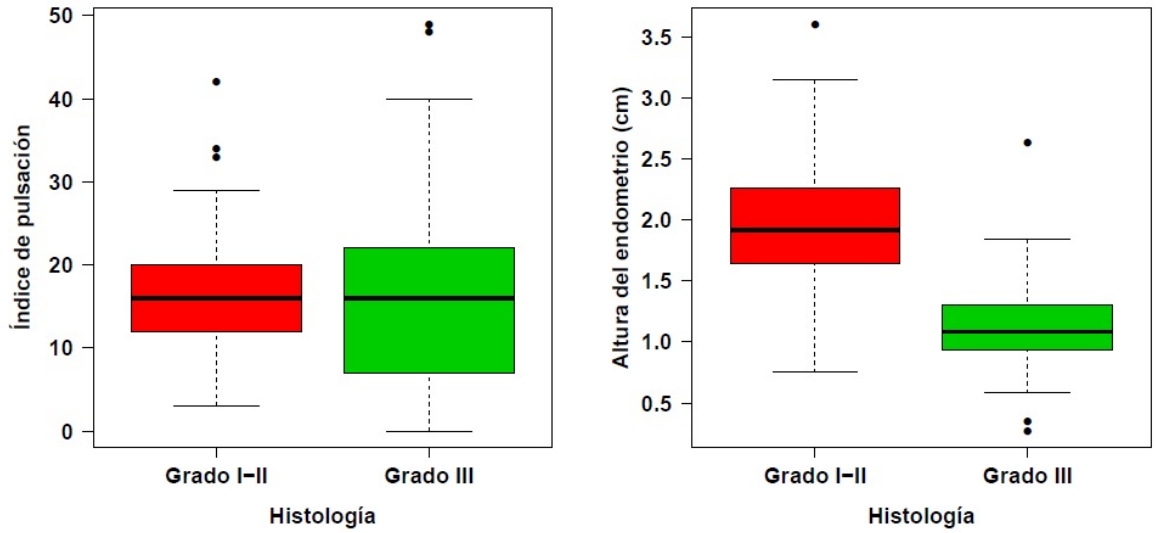


Figura 4.4: Diagramas de cajas de *PI* y *EH* con respecto a *HG*

NV \ *HG*	Grado I–II	Grado III	Total
Ausente	49	17	66
Presente	0	13	13
Total	49	30	79

Tabla 4.3: Tabla de contingencia de *HG* y *NV*

Se muestra cómo no existe separación tanto en *PI* como en *EH*, ya que observamos claramente que los datos se sobreponen en la Figura 4.4. En estas gráficas podemos ver que el grado de histología no varía según el índice de pulsación, pues hay muy poca diferencia entre donde están situadas las cajas, pero en cambio sí que varía según la altura del endometrio, produciéndose mayor grado de histología cuanto menor es la altura del endometrio.

Pasando a *NV*, vemos en la Tabla 4.3 que no hay observaciones si la histología es de grado I–II y la neuvasculización está presente, por lo que se detecta separación cuasicompleta causada por este factor de riesgo. Esto nos llevará a que la estimación de este parámetro a través del método de máxima verosimilitud sea infinita:

$$\widehat{OR} = \frac{49 \cdot 13}{17 \cdot 0} = \frac{637}{0} \longrightarrow \infty$$

Por tanto, aunque a priori hemos visto que existe separación en los datos, ya estamos en disposición de realizar el análisis para poder comparar los $\widehat{OR}$ y sus intervalos de confianza al 95 % por medio del método de máxima verosimilitud y el método de reducción de sesgo. Pero previamente, cabe recordar que hemos descartado utilizar el método de regresión logística exacto porque, como dijimos cuando explicamos estos modelos, uno de los grandes inconveniente que tienen es que puede dar resultados erróneos si alguna de las variables es continua, y en nuestro caso hay dos.

Empezamos definiendo el modelo de regresión logística para este análisis, aunque antes debemos destacar que para la variable *PI* vamos a basar los *OR* en 10 unidades y para la *EH* en 0,1. Esto se debe a que, como vimos en la Tabla 4.2, el rango de *PI* está entre 0 y 49 y el rango de *EH* entre 0,27 y 3,61, lo que puede ocasionar que cada diferencia de unidad provoque cambios muy pequeños para *PI* y muy grandes para *EH* en el análisis. Por este motivo, el modelo para este análisis se define como:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot NV + 10 \cdot \beta_2 \cdot PI + 0,1 \cdot \beta_3 \cdot EH$$

donde p es la probabilidad de que la histología del cáncer sea de grado III según los tres factores de riesgo.

Procedemos a realizar el análisis con R, donde usaremos la función *glm* para el enfoque de máxima verosimilitud (*ML*) y la función *logisticf* [29] para el método de Firth (*FL*). Estas funciones nos devuelven los estimadores $\beta_i, i = 0, 1, 2, 3$, por lo que los convertimos en $\widehat{OR}$ a través de la exponencial, ya que $odds = e^{logit(p)}$. En la Tabla 4.4 presentamos los datos que hemos obtenidos.

Método	Factor de riesgo	$\widehat{OR}$	Intervalos de confianza al 95 %	
ML	*NV*	7,90 $\times 10^7$	0,00	–
	PI	0,66	0,28	1,56
	EH	0,75	0,63	0,88
FL	*NV*	18,71	1,84	2577,65
	PI	0,71	0,29	1,50
	EH	0,77	0,65	0,88

Tabla 4.4: Estimaciones *odds ratio* e intervalos de confianza al 95 % para los tres factores de riesgo

Comparando ambos métodos, observamos que las estimaciones puntuales para *PI* y *EH* mediante el análisis de *FL* están ligeramente más cerca de 1 que por *ML*, tal como cabría esperar por la propiedad de reducción de sesgo del método de Firth. El caso para *NV* es distinto, puesto que si el *OR* estimado por *FL* de 18,71 es un

resultado plausible, todo lo contrario pasa al de *ML*, que es extremadamente grande y teóricamente debería ser infinito.

En cuanto a los intervalos de confianza, vemos que para *FL* los factores *PI* y *EH* son bastante similares que para *ML*, aunque algo más pequeños. No pasa lo mismo para *NV*, donde existe gran diferencia ya que *ML* no da ninguna información, siendo un intervalo unilateral de 0 a infinito.

Por lo tanto, de esta manera queda totalmente demostrado que, cuando está presente el problema de la separación en los datos, utilizando el método de reducción de sesgo se obtiene una mejor información sobre los factores de riesgo que utilizando el método de máxima verosimilitud.

Para acabar este análisis, tenemos que hablar de la interpretación de los parámetros, donde se puede ver claramente que la variable *NV* es un muy fuerte factor de riesgo, algo que ya podíamos intuir por la Tabla 4.3, interpretándose como que tener un cáncer endometrial de grado III está completamente ligado a que la neovasculización está presente en las pacientes.

Respecto a las variables continuas, basándonos en los resultados obtenidos por el método de Firth, para *EH* podemos concluir que por cada 0,1 cm que aumenta la altura del endometrio, el *odds* de tener un cáncer de tipo I–II es $1/0,77 = 1,3$ veces mayor que el de tener uno de tipo III.

Algo parecido pasa con *PI*, puesto que según los resultados se puede intuir que cuanto mayor es el índice de pulsación de la arteria uterina (por cada 10 unidades), mayor será el *odds* de que el cáncer sea de tipo I–II. Pero aquí hay que tener mucho cuidado ya que, a diferencia de *EH*, tenemos que el $\widehat{OR}$ es menor que 1 pero su intervalo de confianza sí que contienen al 1, por lo que no podemos estar seguros de la implicación de *PI* sobre *HG*.

Capítulo 5

Conclusiones

En este trabajo nos hemos centrado en presentar el problema de la separación, un fenómeno que sucede con bastante frecuencia en los modelos de regresión logística, que consiste en la división de las variables respuestas y no respuestas y que es desconocido por muchos investigadores incluso hoy en día.

Este problema tiene gran importancia, ya que una de sus consecuencias más destacable es que impide hallar las estimaciones por medio del método de máxima verosimilitud tradicional. Computacionalmente el problema empeora, porque las limitaciones de los softwares, junto con la mala información de los usuarios, hacen que los resultados que se obtienen sean erróneos y estén muy lejos de la realidad.

Otra cosa que llama la atención es cómo la probabilidad de que se produzca separación suele depender del tamaño de la muestra, del número de factores dicotómicos, de las magnitudes de los *odds ratios* asociados a estos factores y del grado de balanceo de los grupos, por lo que nos lleva a concluir que este problema se presenta en su mayoría en estudios con muestras pequeñas y datos dispersos con varios factores de riesgo altamente predictivos y no balanceados.

A lo largo del trabajo hemos visto posibles soluciones, pero nos hemos focalizado en el método de reducción de sesgo, que fue desarrollado por David Firth, y el cual consiste en modificar la función *score* introduciendo un cierto grado tolerable de sesgo a cambio de la reducción de la variabilidad de las estimaciones de los parámetros. Esto nos conduce a unos estimadores únicos y finitos con iguales o mejores propiedades que los estimadores de máxima verosimilitud.

La superioridad de este método la hemos plasmado a través de un estudio estadístico sobre unos datos de cáncer de endometrio, ayudándonos con el software R y gracias al paquete *logistif*. En este estudio hemos visto la forma en la que se puede detectar si existe o no un posible problema de separación entre alguno de los factores de riesgo y la enfermedad. Debido a que hemos encontrado separación en uno de los factores, hemos pasado a realizar un análisis estadístico mediante el método de máxima verosimilitud y el método de reducción de sesgo para calcular los *odds ratios* estimados para cada factor y sus correspondientes intervalos de confianza al 95 %.

Comparando ambos métodos, hemos demostrado con claridad cómo los resultados del método de Firth dan mucha mejor información que los otros, que incluso llegan a carecer de sentido.

Antes de acabar, cabe destacar que el bioestadístico Georg Heinze presentó, en uno de los seminarios del departamento de Estadística e Investigación Operativa de la UPC, un estudio sobre la regresión logística con eventos raros, donde propuso el método FLAC–FLIC (*Firth's Logistic regression with Added Covariate – Firth type Logistic regression with Intercept Correction*), que no es más que una extensión del método original de Firth donde se le añade a su vez un pequeño factor a los propios datos.

Este método, en principio, fue solo una propuesta y aún carece de ningún artículo donde basarse. Únicamente disponemos de las diapositivas con las que se ayudó Heinze en su presentación [30]. Pero es un ejemplo de cómo el método de Firth, aunque lleve ya algunos años en funcionamiento, no es una técnica definitiva.

Por último, consideramos interesante de cara a futuros trabajos explorar el comportamiento del problema de la separación en los modelos de regresión log-binomial, los cuales se pueden definir como:

$$\ln \mathrm{P}(Y = 1|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$$

donde β_0 es la constante y $\beta_1, \ldots, \beta_m$ son los parámetros del modelo. Modelando la probabilidad de la variable respuesta ($Y = 1$) llegamos a:

$$\mathrm{P}(Y = 1|\boldsymbol{X}) = \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m)$$

Estos modelos son de gran utilidad cuando queremos estimar la razón de prevalencias o el riesgo relativo en un estudio con resultados comunes en la población.

Tras una búsqueda bibliográfica, apenas encontramos información sobre la regresión log-binomial. Y es por ello que su implementación no está lo suficientemente desarrollada en ninguno de los softwares estadísticos.

Bibliografía

[1] K. Langohr (2017). *Apuntes de la asignatura de Epidemiología del Máster en Estadística e Investigacíon Operativa*, UPC–UB.

[2] N. P. Jewell (2004). *Statistics for epidemiology*, Chapman & Hall/CRC.

[3] M. Porta (2008). *A Dictionary of Epidemiology*, 5^{th} ed. Oxford University Press.

[4] M. Hernández Avila, F. Garrido Latorre y S. López Moreno (2000). *Diseño de estudios epidemiológicos*, Salud pública de Mexico, 42, 144–154.

[5] J. Zhang y K. Yu (1998). *What's relative risk? A method of correcting the odds ratio in cohort studies of common outcomes*, Journal of American Medical Association, 280, 1690–1691.

[6] S. Parodi y B. Ezio (2007): *The Mantel-Haenszel Procedure in Epidemiological Studies: An Introduction.* Annali della Facoltá de Medicina Veterinaria di Parma, 17, 17–32.

[7] D. Hosmer, T. Hosmer, S. Le Cessie y S. Lemeshow (1997). *A Comparison of Goodness-of-fit Tests for the Logistic Regression Model*, Statistics in Medicine, 16, 965–980.

[8] D. Hosmer and S. Lemeshow (2000). *Applied Logistic Regression*, 2^{nd} ed. New York: John Wiley & Sons.

[9] S. R. Cole, H. Chu y S. Greenland (2013). *Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer*, American Journal of Epidemiology, 179, 252–260.

[10] J. A. Anderson y V. Blair (1982). *Penalized maximum likelihood estimation in logistic regression and discrimination*, Biometrika, 69, 123–136.

[11] A. Agresti (2011). *Score and pseudo-score confidence intervals for categorical data analysis*, Statistics in Biopharmaceutical Research, 3, 163–172.

[12] D. Firth (1993). *Bias reduction of maximum likelihood estimates*, Biometrika, 80, 27–38.

[13] G. Heinze y M. Shemper (2002). *A solution to the problem of separation in logistic regression*, Statistics in Medicine, 21, 2409–2419.

[14] J. C. Correa y M. Valencia (2011). *La separación en regresión logística, una solución y aplicación*, Revista Facultad Nacional de Salud Pública, 3, 281–288.

[15] M. A. Mansournia, A. Geroldinger, S. Greenland y G. Heinze (2017). *Separation in logistic regression: Causes, consequences and control*, American Journal of Epidemiology, 187, 864–870.

[16] A. Albert y J. A. Anderson (1984). *On the existence of maximum likelihood estimates in logistic regression models*, Biometrika, 71, 1–10.

[17] T. J. Santner y D. E. Duffy (1986). *A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models*, Biometrika, 73, 755–758.

[18] A. Christmann y P. J. Rousseeuw (2003). *Robustness against separation and outliers in logistic regression*, Computational Statistics and Data Analysis, 43, 315–332.

[19] C. R. Mehta y N. R. Patel (1995). *Exact logistic regression: theory and examples*, Statistics in Medicine, 14, 2143–2160.

[20] E. King y T. P. Ryan (2002). *A preliminary investigation of maximum likelihood logistic regression versus Exact logisic Regression*, American Statistical Association, 56, 163–170.

[21] K. F. Hirji, C. R. Mehta y N. R. Patel (1987). *Computing distributions for exact logistic regression*, American Statistical Association, 82, 1110–1117.

[22] D. Firth (1992). *Generalized linear models and Jeffreys priors: an iterative weighted least-squares approach*, In Computational Statistics, 1, Physica-Verlag: Heidelberg, 553–557.

[23] R. L. Schaefer (1983). *Bias correction in maximum likelihood logistic regression*, Statistics in Medicine, 2, 71–78.

[24] AECC, *Asociación Española Contra el Cáncer*. Disponible en: https://www.aecc.es/es/todo-sobre-cancer/tipos-cancer/cancer-utero/que-es-cancer-utero [Último acceso: 2018].

[25] ACM, *American Cancer Society*. Disponible en: https://www.cancer.org/es/cancer/cancer-de-endometrio/acerca/que-es-cancer-de-endometrio [Último acceso: 2018].

[26] NIH, *National Cancer Institute*. Disponible en: https://www.cancer.gov/espanol/tipos/uterino/pro/tratamiento-endometrio-pdq [Último acceso: 2018].

[27] I. Marschner y M. W. Donoghoe. *glm: Fitting Generalized Linear Models*. Versión 1.2.1, 2018. https://cran.r-project.org/web/packages/glm/

[28] I. Kosmidis, K. Konis, E. C. Kenne Pagui y N. Sartori. *brglm2: Bias Reduction in Generalized Linear Models*, Versión 0.1.8, 2018. https://cran.r-project.org/web/packages/brglm2/

[29] G. Heinze, M. Ploner, D. Dunkler y H. Southworth. *logistf: Firth's Bias-Reduced Logistic Regression*, Versión 1.23, 2018. https://cran.r-project.org/web/packages/logistf/

[30] G. Heinze (2017). *Losgistic regression with rare events: problems and solutions*. Diapositivas disponibles en: https://www.eio.upc.edu/ca/seminari/docs/georg-heinze-logistics-regression-with-rare-events-problems-and-solutions.pdf

Apéndice A

Código R

```
#########################################################################
################    Estudio de cancer de endometrio    ################
################  Estimadores de maxima verosimilitud  ################
################         Estimadores de Firth          ################
#########################################################################

#----------------------------
#__________Paquetes__________
#----------------------------

install.packages('Epi')
library(Epi)
#----------------------------
install.packages('brglm2')
library(brglm2)
#----------------------------
install.packages('descr')
library(descr)
#----------------------------
install.packages('beeswarm')
library(beeswarm)
#----------------------------
install.packages('logistf')
library(logistf)

#-----------------------------------------
#____________Conjunto de datos____________
#-----------------------------------------
```

```
# Prepararacion de los datos:
?endometrial
data(endometrial)
View(endometrial)
summary(endometrial)
mode(endometrial)

# Cambios y comentarios en los datos:
endometrial$HG <- factor(endometrial$HG,
                  labels = c("Grado I-II", "Grado III"))
endometrial$NV <- factor(endometrial$NV,
                  labels = c("Ausente", "Presente"))
comment(endometrial) <- "Datos del grado de histologia y factores de
                        riesgo de 79 mujeres con cancer de endometrio"
comment(endometrial$HG) <- "Histologia del endometrio: 0 si grado I-II,
                            1 si grado III"
comment(endometrial$NV) <- "Ausencia (0) o presencia (0) de
                            neurovasculizacion"
comment(endometrial$PI) <- "Indice de pulsacion de la arteria uterina"
comment(endometrial$EH) <- "Altura del endometrio (cm)"
str(endometrial)

# Frecuencias de HG y NV:
freq(endometrial$HG, plot = F)
freq(endometrial$NV, plot = F)

# Resumen descriptivo de PI y EH:
summary(endometrial$PI, endometrial$EH)

# Diagramas de cajas de PI y EH con respecto a HG:
nums <- which(sapply(endometrial, is.numeric))
for (i in nums) {
   pdf(width = 12, height = 6)
   par(las = 1, mfrow = c(1,2), font.lab = 2, font.axis=2, oma =
       c(0,0,1,0), mar = c(5,5,2,2), cex.axis = 1.3, cex.lab = 1.3)
   boxplot(PI~HG, endometrial, col = 2:3, pch = 16, xlab =
           "Histologia", ylab = "Indice de pulsacion")
   boxplot(EH~HG, endometrial, col = 2:3, pch = 16, xlab =
          "Histologia", ylab = "Altura del endometrio (cm)")
```

```
   dev.off()
}

# Tabla de contiengencia de HG y NV:
tabfrec <- table(endometrial$NV, endometrial$HG)
names(dimnames(tabfrec)) <- c("Neurovasculizacion", "Histologia")
tabfrec


#-----------------------------------------------------
#___________Metodo de maxima verosilimitud___________
#-----------------------------------------------------

# Regresion logistica con "glm":
ML <- glm(HG ~ NV + PI + EH, endometrial, family = 'binomial')
summary(ML)

# Variable NV con respecto a HG:
ML_NV <- round(cbind(exp(ML$coef[2]),
         exp(summary(ML)$coef[2,1] - 1.96 * summary(ML)$coef[2,2]),
         exp(summary(ML)$coef[2,1] + 1.96 * summary(ML)$coef[2,2])), 2)
# Variable PI con respecto a HG:
ML_PI <- round(cbind(exp(ML$coef[3] * 10),
         exp(summary(ML)$coef[3,1] * 10 - 1.96 * summary(ML)$coef[3,2]
         * 10),
         exp(summary(ML)$coef[3,1] * 10 + 1.96 * summary(ML)$coef[3,2]
         * 10)), 2)
# Variable EH con respecto a HG:
ML_EH <- round(cbind(exp(ML$coef[4] * 0.1),
         exp(summary(ML)$coef[4,1] * 0.1 - 1.96 * summary(ML)$coef[4,2]
         * 0.1),
         exp(summary(ML)$coef[4,1] * 0.1 + 1.96 * summary(ML)$coef[4,2]
         * 0.1)), 2)

# Tabla de odds ratio estimados e intervalos de confianza:
ML_resultados <- rbind(ML_NV, ML_PI, ML_EH)
colnames(ML_resultados) <- c("OR", "IC_inf", "IC_sup")
ML_resultados
```

```
#-----------------------------------------------------
#_____________Metodo de reduccion de sesgo_____________
#-----------------------------------------------------

# Regresion logistica con "logistf":
FL <- logistf(HG ~ NV + PI + EH, data = endometrial)
summary(FL)

# Variable NV con respecto a HG:
FL_NV <- round(cbind(exp(FL$coef[2]),
         exp(FL$ci.lower[2]), exp(FL$ci.upper[2])), 2)
# Variable PI con respecto a HG:
FL_PI <- round(cbind(exp(10 * FL$coef[3]),
         exp(10 * FL$ci.lower[3]), exp(10 * FL$ci.upper[3])), 2)
# Variable EH con respecto a HG:
FL_EH <- round(cbind(exp(0.1 * FL$coef[4]),
         exp(0.1 * FL$ci.lower[4]), exp(0.1 * FL$ci.upper[4])), 2)

# Tabla de odds ratio estimados e intervalos de confianza:
FL_resultados <- rbind(FL_NV, FL_PI, FL_EH)
colnames(FL_resultados) <- c("OR", "IC_inf","IC_sup")
FL_resultados
```